wCorr Proofs

Paul Bailey

2016-11-21

Proof of consistency of Horvitz-Thompson estimator of a mean.

An Horvitz-Thompson (HT) estimator of a sum takes the form

$$\hat{Y} = \sum_{i=1}^{n} \frac{1}{\pi_i} y_i \ , \tag{1}$$

where there are n sampled units from a population of N units, each unit has a value $y \in \Re$, each unit is sampled with probability π_i, and $\hat{Y}$ is the estimated of Y in a population. Probabilisticaly, the formula is

$$\hat{Y} = \sum_{i=1}^{N} I_i \frac{1}{\pi_i} y_i \ , \tag{2}$$

where I_i is one when a unit sampled and zero otherwise. Notice that

$$E\left(I_i\right) = \pi_i \ , \tag{3}$$

and that the covariance of two units ($\mathrm{Cov}\left(I_i, I_j\right)$ for $i \neq j$) is not assumed to have any particular structure.

Unbiasedness of Horvitz-Thompson estimator

Given:

1. $\frac{n}{N} = v$, for some $v \in \Re$,
2. $\frac{1}{\pi_i} > 0 \forall i \in \{1, ..., N\}$

It is simple to see that

$$\begin{aligned} E(\hat{Y}) &= \sum_{i=1}^{N} \frac{E\left(I_i\right)}{\pi_i} y_i & (4)\\ &= \sum_{i=1}^{N} y_i & (5)\\ &= Y & (6) \end{aligned}$$

Decreasing variance of the Horvitz-Thompson estimator

The the variance of the HT estimator is, switching to the now more conveniant notation where $w_i = \frac{1}{\pi_i}$

$$\mathrm{Var}(\hat{Y}) = \mathrm{Var}\left(\sum_{i=1}^{N} I_i w_i y_i\right) \tag{7}$$

$$= \sum_{i=1}^{N} \mathrm{Var}\left(I_i\right) w_i^2 y_i^2 + \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} w_i y_i w_j y_j \mathrm{Cov}\left(I_i, I_j\right) \ . \tag{8}$$

Because the I_i has a Bernouli distribution, $\mathrm{Var}\left(I_i\right) \leq 1/4$. To bound the covariance terms, consider building the variance of the sum elementwise. First, define

$$A = \sum_{i=1}^{m-1} y_i \tag{9}$$

$$B = y_m \tag{10}$$

then

$$\mathrm{Var}\left[\left(\sum_{i=1}^{m-1} y_i\right) + y_m\right] = \mathrm{Var}\left(A + B\right) \tag{11}$$

$$= \sigma_A^2 + \sigma_B^2 + 2\sigma_{AB} \ , \tag{12}$$

where σ_A^2 is the variance of A, σ_B^2 is the variance of B, and σ_{AB} is the covariance of A and B. The formula that relates the correlation coefficient ρ to the standard deviations and covaraince is

$$\rho = \frac{\sigma_{AB}}{\sigma_A \sigma_B} \tag{13}$$

$$\sigma_A \sigma_B = \frac{\sigma_{AB}}{\rho} \tag{14}$$

this is clearly maximized when ρ is one in magnitude and has the same sign as σ_{AB} so that

$$\sigma_A \sigma_B \leq |\sigma_{AB}| \tag{15}$$

plugging that into (12) results in

$$\mathrm{Var}\left(A + B\right) \leq \sigma_A^2 + \sigma_B^2 + 2\sigma_A \sigma_B \ , \tag{16}$$

define $k \in \Re$ such that

$$\sigma_B = k\sigma_A \tag{17}$$

then (**??**) becomes

$$\begin{aligned} \mathrm{Var}\,(A+B) &\leq \sigma_A^2 + k^2\sigma_A^2 + 2k\sigma_A^2 \ , && (18) \\ &= \sigma_A^2 \left(1 + 2k + k^2\right) \ . && (19) \end{aligned}$$